SMEROVANIE VÝVOJA JAZYKOVÝCH NÁSTROJOV A TECHNOLÓGIÍ

Adriana Žáková
Slovenský národný korpus JÚĽŠ SAV, Panská 26, 813 64 Bratislava
e-mail: adriana163@korpus.sk

V priestoroch hotela Park Inn Danube v Bratislave sa v dňoch 7. a 8. júna 2012 konala medzinárodná konferencia Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu)**,** ktorá sa realizovala pri príležitosti 10. výročia vzniku špecializovaného pracoviska Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra Slovenskej akadémie vied. Prvý deň podujatia bol zároveň jedným z prezentačných dní členských krajín projektu CESAR, ktorého cieľom je sprístupniť existujúce jazykové zdroje a prispieť k tvorbe jednotného digitálneho trhu v Európe. Prezentačné dni sú zamerané na spoločné aktivizovanie národného úsilia o rozvoj a výskum jazykových technológií. Slovenskému prezentačnému dňu predchádzalo podobné podujatie v máji tohto roku v Bulharsku, ktoré je spolu so Slovenskom, Chorvátskom, Poľskom a Srbskom zapojené do projektu CESAR koordinovaného partnerom z Maďarskej akadémie vied. Hlavným cieľom bratislavskej konferencie bolo predstaviť jazykové technológie a nástroje počítačového spracovania prirodzeného jazyka a zosumarizovať súčasný stav jazykových technológií na Slovensku a v okolitých krajinách.

Pred otvorením samotného vedecko-informačného podujatia sa konala tlačová konferencia za účasti podpredsedu SAV Ľubomíra Falťana, generálnej riaditeľky sekcie vedy a techniky MŠVVaŠ SR Marty Cimbákovej, koordinátora projektu CESAR Tamása Váradiho, riaditeľa JÚĽŠ SAV Pavla Žiga, vedúcej oddelenia Slovenského národného korpusu JÚĽŠ SAV Márie Šimkovej a hlavného riešiteľa slovenskej časti projektu CESAR Radovana Garabíka. Pozvaní účastníci tlačovej konferencie stručne predstavili projekt CESAR, účasť Slovenska v ňom a význam takýchto projektov pre malé jazyky.

Úvodné slovo nasledujúceho slávnostného otvorenia patrilo Ľ. Falťanovi, ktorý poukázal na dôležitosť jazyka v každodennej komunikácii a v mene SAV vyjadril plnú podporu oddeleniu SNK JÚĽŠ SAV. M. Cimbáková vo svojom príhovore vyzdvihla moderné spracovanie prirodzeného jazyka s neprehliadnuteľným významom v akademickej, výskumnej i verejnej sfére. P. Žigo zdôraznil prínos moderných jazykových technológií, vďaka ktorým sú výsledky výskumu oddelenia SNK dostupné pre širokú verejnosť.

Prvý blok rokovania otvoril Georg Rehm z Nemeckého výskumného centra umelej existencie v Berlíne, ktorý vo svojom príspevku Projekt META-NET zdôraznil potrebu podpory technologických základov multilingválnej európskej informačnej spoločnosti. V prezentácii autor predstavil aj sériu bielych kníh, ktoré opisujú stav jazykových zdrojov v jednotlivých krajinách Európy. Zo zhrnutých výsledkov slovenskej bielej knihy

vyplynulo, že slovenčina má v oblasti rečovej a textovej analýzy len veľmi slabú podporu, pričom niektoré oblasti, napr. strojový preklad, nie sú takmer vôbec pokryté. Na záver svojej prezentácie autor predstavil Strategickú výskumnú agendu, ktorá zoskupuje odborníkov z oblasti prekladu, lokalizácie, informačných služieb a interaktívnych systémov. Ich úlohou je vytvoriť spoločnú víziu budovania komunity zaoberajúcej sa jazykovými technológiami v Európe.

Projekt CESAR bol predmetom príspevku Tamása Váradiho z Maďarskej akadémie vied, ktorý stručne predstavil cieľ spoločne zdieľať digitálne zdroje. Autor vyzval všetky zainteresované strany – národný priemysel, výskumníkov, štátne inštitúcie a i., aby podporovali budovanie a zdokonaľovanie jazykových technológií. T. Váradi vyzdvihol veľkosť slovenského národného korpusu (770 miliónov tokenov), no zároveň konštatoval, že na slovenskom trhu chýbajú sémantické korpusy a technológie na parsovanie viet.

Slovenský národný korpus (2002 – 2012): východiská, ciele a výsledky pre výskum a prax prezentovali Mária Šimková a Radovan Garabík z oddelenia SNK JÚĽŠ SAV. M. Šimková uviedla dôvody vzniku (najmä potrebu dostatočnej materiálovej bázy pre lexikografické opisy slovenčiny) a zhrnula priebeh budovania oddelenia SNK, ktoré súviselo aj s celkovým rozmachom korpusovej lingvistiky vo svete. Už prvotný interný korpus so základným počítačovým vybavením sa JÚĽŠ SAV snažil podporiť medzinárodnými projektmi. O potrebe a rozšírení SNK svedčí aj približne 500 registrovaných používateľov korpusu ročne a priemerne 40 000 dopytov denne do databázy lingvistických zdrojov. O medzinárodných projektoch oddelenia informoval R. Garabík. Okrem projektov CESAR a META-NET predstavil aj projekty Mondilex, Slovak Online, EuroMatrixPlus a niekoľko spoluprác v rámci medziakademických dohôd. Vďaka realizácii projektov a finančnej podpore Európskej komisie sa vybudoval 5-jazyčný slovník sémantických vzťahov, morfologická databáza, webový korpus, špecializovaný korpus právnych textov, slovenský hovorený korpus a niekoľko paralelných korpusov.

Dopoludňajšiu sekciu uzavrel Marko Tadić z Filozofickej fakulty Univerzity v Záhrebe príspevkom Chorvátsky národný korpus a jeho úloha pri budovaní jazykových technológií v Chorvátsku. Autor vo svojej prezentácii ponúkol prierez všetkých verzií Chorvátskeho národného korpusu (HNK) datovaných od roku 1967 až po súčasnosť. Spomenul problémy s nedostatkom elektronických textov a chýbajúcimi prepismi v začiatkoch budovania korpusu. Od roku 2005 je HNK lematizovaný s možnosťou rozšíreného vyhľadávania prostredníctvom korpusového manažéra Manatee s klientom Bonito. Autor zdôraznil, že na rozširovanie korpusu je potrebná riadená podpora výskumu a rozvoja zdrojov a nástrojov. M. Tadić vyjadril aj obdiv nad tým, že sa SNK vybudoval v kratšom čase ako HNK a v súčasnosti ho už v počte tokenov presahuje.

Vstupom do popoludňajšej časti prvého rokovacieho dňa boli prezentácie slovenských a českých firiem z oblasti jazykových a informačných technológií. Ako

prvý vystúpil Vladimír Kadlec z firmy Seznam.cz. Autor uviedol štatistiky fulltextového vyhľadávania a zaindexovaných stránok (380 miliónov českých, 130 miliónov anglických, 15 miliónov slovenských atď.). Seznam.cz označil za otvorenú firmu, ktorá prijíma nové dáta a nápady. Spoločnosť NEWTON Technologies predstavili Petr Herian a Pavel Barták. V prezentácii hovorili o systéme automatického prepisu reči NEWTON Dictate s 90 % úspešnosťou presného prepisu. Autori názorne ukázali prácu so systémom. Pre slovenský trh majú systém okrem všeobecného slovenského lexikónu natrénovaný na oblasť justície, všeobecnej medicíny, rádiológie a patológie. Peter Baláž predstavil neziskovú organizáciu Edukácia@Internet, ktorá sa angažuje v oblasti medzikultúrneho vzdelávania a používania jazykov s podporou jazykových technológií. Autor predstavil najväčší on-line portál na výučbu esperanta na svete (www.lernu.net), ktorý má v súčasnosti 117 000 používateľov z 33 krajín. Organizácia vytvorila aj viacjazyčný portál Slovak Online na výučbu slovenského jazyka, na ktorom participovali aj pracovníci SNK JÚĽŠ SAV.

Po prezentáciách firiem, ktoré boli okrem týchto vstupov k dispozícii aj v posterovej sekcii, nasledovala panelová diskusia o postavení slovenského jazyka v digitálnom veku. Účastníci diskusie sa vyjadrili o potrebe širšieho sprostredkúvania jazykových dát a modelov z vedeckého prostredia do aplikačného a komerčného využitia, poukázali na slabú finančnú podporu národných vlád projektom v tejto oblasti a vyzdvihli nevyhnutnosť zdieľania zdrojov a dôležitosť dostupnosti digitalizovaného textu pre budúce generácie.

Dušan Katuščák zo Slovenskej národnej knižnice v Martine sa publiku prihovoril prostredníctvom príspevku Národný projekt SNK Digitálna knižnica a digitálny archív, ktorého cieľom je digitalizovať pôvodné slovacikálne dokumenty, teda slovenské písomné kultúrne dedičstvo. Po vybudovaní infraštruktúry pre digitalizačnú a konzervačnú činnosť by sa Slovensko zaradilo medzi európskych lídrov v oblasti digitalizácie. Jan Hajič z Matematicko-fyzikálnej fakulty Karlovej univerzity v Prahe prezentoval projekt LINDAT-Clarin, jazyková infraštruktúra pre výskum, ktorý sa zameriava na zber jazykových dát z oblasti humanitných vied. Autor vyzýval k zdieľaniu i menších jazykových zdrojov. Veľmi dôležitou časťou projektu je podpora výskumníkov a študentov prostredníctvom letných škôl a iných tréningov. Riaditeľ Ústavu informatiky SAV v Bratislave Ladislav Hluchý vystúpil s témou Počítačové technológie na spracovanie reči a textu. V prezentácii spomenul vytváranie vysokovýkonnej infraštruktúry na základe klastrového a gridového počítania pre spoločenské a humanitné vedy. Ako posledná odznela v prvý deň rokovania prezentácia Praktické aplikácie automatického spracovania reči na ÚI SAV od Milana Ruska z Oddelenia spracovania reči ÚI SAV v Bratislave. Autor predstavil webový slovník gest DiGest – Dictionary of Gestures, vyučovací systém na kontrolu výslovnosti EURONOUNCE a systémy na syntézu a rozpoznávanie reči.

Druhý deň konferencie otvoril František Čermák, riaditeľ Ústavu Českého národného korpusu Filozofickej fakulty Karlovej univerzity v Prahe, príspevkom Projekt InterCorp a jeho povaha. Autor niekoľkokrát vyzdvihol potrebu budovať korpusy na účely štúdia jazykových javov v kontexte. Projekt InterCorp buduje paralelné korpusy pre jazyky, ktoré sa študujú na FF UK v Prahe. On-line paralelné korpusy slúžia výskumníkom i študentom ako bohatý zdroj informácií pre teoretický výskum, výučbu cudzích jazykov, dvoj- a viacjazyčnú lexikografiu a pod. Ďalší účastníci z Českej republiky Karel Pala a Pavel Rychlý z Fakulty informatiky Masarykovej univerzity v Brne v príspevku Budovanie veľkých korpusov a nástrojov pre počítačovú lexikografiu predstavili niekoľko korpusových nástrojov: WebBootCat – na vyhľadávanie konkrétnych domén na webe podľa kľúčového slova, SpiderLing – na získavanie textov, JusText – na odstraňovanie netextových častí webových stránok, onion – na odstraňovanie dátových duplicít. Pre lexikografov autori ponúkli nástroje: DEB – editor a prehliadač slovníkov, DEBDict – všeobecný prehliadač hlavných českých slovníkov, DEBVisDic – editor a prehliadač sémantických sietí, TeDi – na budovanie terminologických slovníkov, PDEV – na spájanie sémantiky slova s použitím v texte. Viacjazyčné zdroje pre bulharčinu – najnovší vývoj (skúsenosti IMI BAS) prezentovala Ludmila Dimitrova z Bulharskej akadémie vied. Autorka ponúkla prehľad niekoľkých projektov a medziakademických dohôd so zameraním na budovanie paralelných korpusov, vďaka ktorým sa podarilo vybudovať aj slovensko-bulharský paralelný korpus s veľkosťou 1,2 milióna slov. Zástupca Ruskej akadémie vied Leonin Iomdin prišiel na podujatie s prezentáciou Automatizované spracovanie textu a hĺbkovo syntakticky anotovaný korpus ruských textov: ich interakcia a vzájomný vplyv. V úvode autor predstavil typy Ruského národného korpusu a syntaktický podkorpus SynTagRus, ktorý obsahuje 52 000 viet rôznych žánrov. Základ podkorpusu tvorí závislostná anotačná schéma Pražského závislostného korpusu. L. Iomdin uviedol, že SynTagRus dokáže rozpoznať 75 syntaktických vzťahov a na morfologické značkovanie využíva Ruský morfologický slovník, ktorý obsahuje 130 000 hesiel.

V popoludňajšom bloku Pavol Žigo, riaditeľ JÚĽŠ SAV, predstavil v prezentácii Počítačová podpora kartografického spracovania nárečí slovanských jazykov dialektologický korpus, ktorý mapuje 3 454 nárečových javov z 850 slovanských lokalít od Stredozemného mora po Ural. Autor demonštroval výsledky bádania na pracovnej mape. Uviedol napríklad fakt, že slovanské jazyky hraničiace s neslovanskými prestali postupom času skloňovať a funkciu koncoviek v nich prebrali predložky (bulharčina, macedónčina). Július Kravjar z Centra vedecko-technických informácií SR v Bratislave prezentoval atraktívnu tému Národný korpus záverečných prác slovenských vysokých škôl a boj proti plagiátorstvu. Od roku 2008 sa na Slovensku všetky záverečné práce zhromažďujú v spoločnom centrálnom registri. Spolu s antiplagiátorským systémom ho vytvorila spoločnosť SVOP. Unikátnosť projektu spočíva v tom, že sa reali-

zuje celoplošne na území Slovenska. V registri sa aktuálne nachádza 214 000 prác. Po dvoch rokoch realizácie projektu J. Kravjar skonštatoval, že sa zvýšila kvalita prác, ako aj povedomie o autorských právach. Tibor Pintér z Maďarskej akadémie vied orientoval svoje rozprávanie na maďarský národný korpus a jeho rozšírenie v prezentácii MAĎARSKÝ NÁRODNÝ KORPUS II – POKUS O VEĽKÝ GIGABAJTOVÝ KORPUS. Autor sa vyjadril, že snahou tvorcov nového národného korpusu je vytvoriť miliardový korpus so štýlovo i žánrovo pestrými textami, ktorý by slúžil ako zdroj pri koncipovaní jazykových derivátov, ako napr. n-gramov, frekvenčných slovníkov a pod.

Druhú časť popoludňajšej sekcie otvorili výskumníci Ján Staš, Daniel Hládek a Jozef Juhár z Technickej univerzity v Košiciach, ktorí prišli na podujatie s témou BUDOVANIE ORGANIZOVANÉHO KORPUSU TEXTOV PRE REČOVÉ TECHNOLÓGIE V SLOVENČINE. Názorne predviedli systém automatického prepisu reči, ktorý dokáže rozoznávať okrem jednoslovných tokenov aj skloňované skratky, akronymá, regulárne výrazy a i. Ich systém CisloSlovom dokáže podľa kontextu správne zapísať *číslo dva*, napríklad: *dvaja* ľudia, *2.* apríla, zaplatili *2* koruny. Autori sa vo svojom výskume zameriavajú najmä na oblasť justície, pre ktorú vytvorili jazykový model s 95 % úspešnosťou. Odborné vystúpenia podujatia uzavrela Velislava Stojkova z Bulharskej akadémie vied prezentáciou KOLABORATÍVNE VYVINUTÉ LEXIKÁLNE ZDROJE PRE BULHARČINU S APLIKÁCIOU NA TVORBU SLOVNÍKOV A REFERENČNÝCH ZDROJOV. Autorka sa primárne venovala výskumu extrakcie sémantických vzťahov medzi pojmami v stredoškolskej matematike. Pracuje s korpusmi anglických a bulharských matematických textov MathWiki a MathWiki-Bul, pričom využíva štatistický prístup vyhľadávania najfrekventovanejších pojmov prostredníctvom nástroja Sketch Engine. Po zrealizovaní výskumu možno označiť za hyperonymum pojem *complex function*, ktorého hyponymami sú napr. *polynomial function* alebo *exponential function*. V. Stojkova označila metódu štatistického prístupu za rýchly a moderný spôsob pri tvorbe referenčných zdrojov.

Záverečné slovo patrilo vedúcej oddelenia Slovenského národného korpusu JÚĽŠ SAV M. Šimkovej a riaditeľovi Jazykovedného ústavu Ľ. Štúra SAV P. Žigovi, ktorí sa poďakovali všetkým účastníkom za príjemnú rokovaciu atmosféru i prínosné diskusie, organizátorom za skvelý priebeh podujatia a zapriali oddeleniu Slovenského národného korpusu ešte veľa úspešných rokov pri rozširovaní a budovaní nových korpusov a korpusových nástrojov.

Medzinárodná konferencia Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu) bola príležitosťou, na ktorej si 21 prednášajúcich i 99 prítomných hostí z rozmanitých vedeckých, akademických i komerčných oblastí vymenili najnovšie poznatky z oblasti spracovania prirodzeného jazyka a zároveň vyjadrili svoje predstavy o budúcom smerovaní vývoja jazykových nástrojov a technológií. Prednesené príspevky budú publikované v zborníku Jazykovedné štúdie XXXI.